Using Differential Privacy to Protect Earnings and Employment Statistics

Andrew Foote
Principal Economist, Center for Economic Studies
U.S. Census Bureau

National Advisory Committee

October 27, 2022

What are Post-Secondary Employment Outcomes and Veteran Employment Outcomes?

Post-Secondary Employment Outcomes

- Partnership between Census Bureau and university systems or state longitudinal data systems
- Provides national earnings and employment outcomes by institution, degree and field

Veterans Employment Outcomes

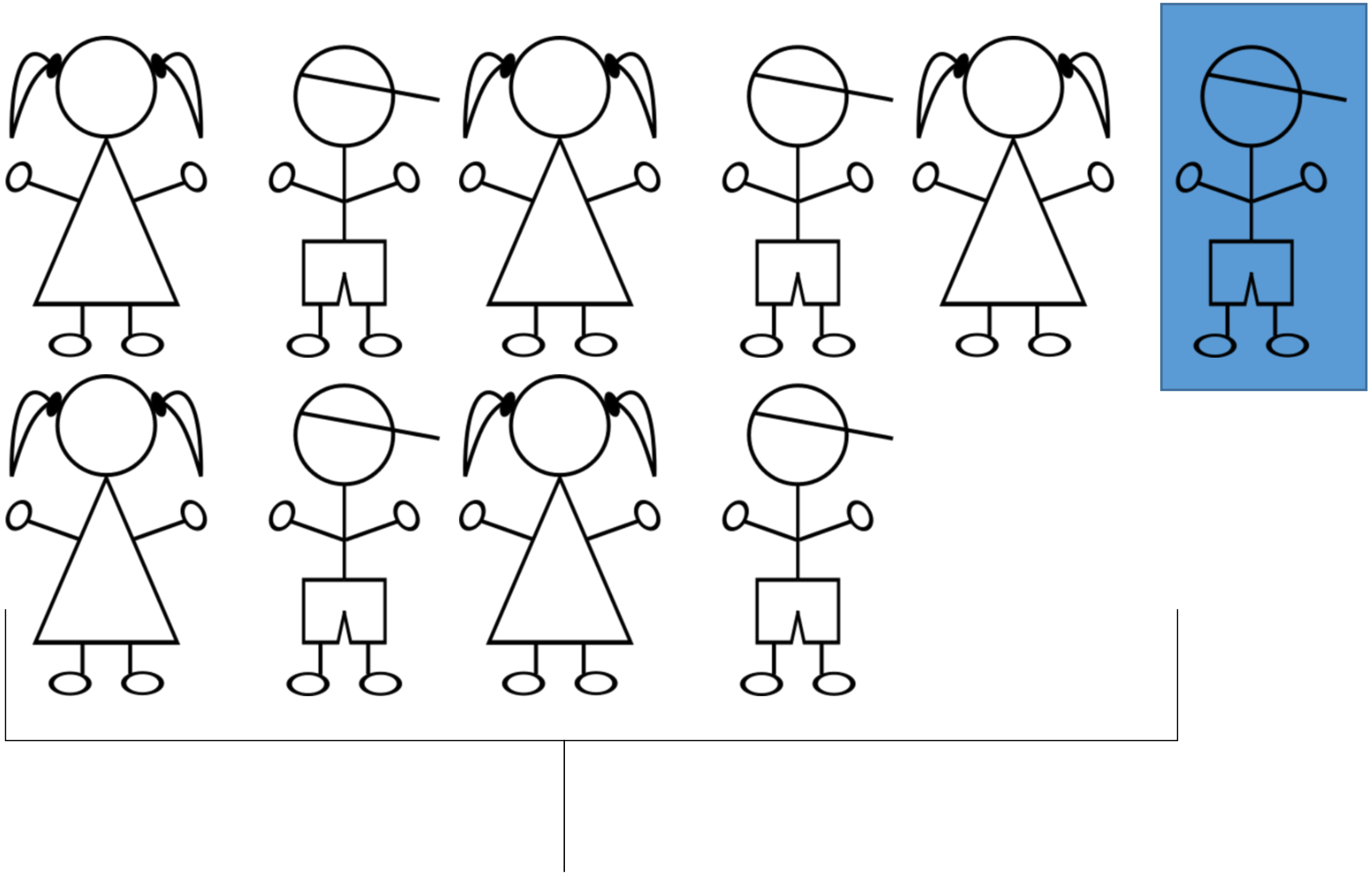
- Partnership between Census and Army
- Provides national earnings and employment outcomes military specialization, service characteristics, employer industry (if employed), and veteran demographics

United States® Census Bureau

What are the threats to student privacy?

Protecting the Microdata

- Title 13 requirement:
  - The existence of a job held by an individual is confidential
- **We do not have a monopoly on microdata** – most of our partners have access to the frame (all graduates) and most of the earnings data we use to produce the statistics



In-State Median Earnings: $80,000

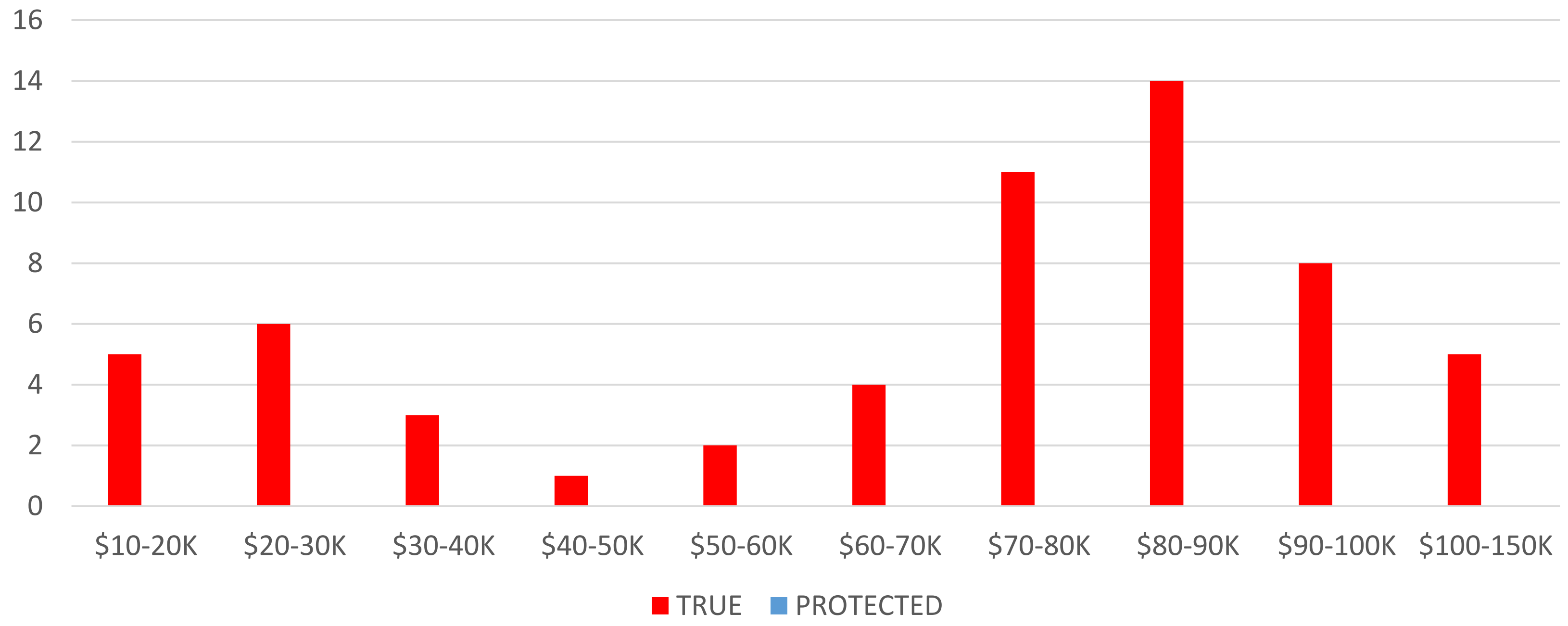
National Median Earnings: $85,000

How we implement Differential Privacy

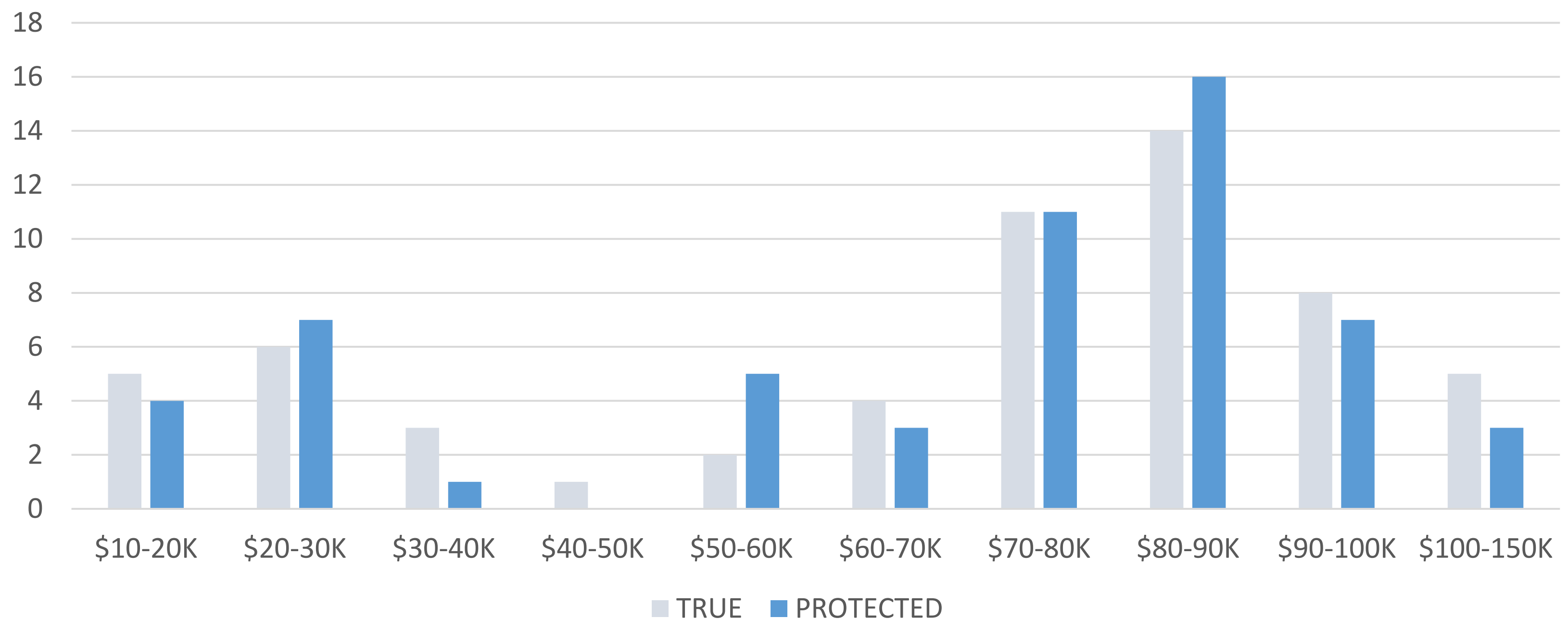
- Construct a histogram of earnings for each cell
  - Log-normal with parameters based on ACS public-use sample
  - Histogram bin ranges are public information
- Add geometric noise to each histogram bin
- Extract percentiles from the resulting protected CDF
- We use a total privacy loss budget of 3, half of which is used on the earnings tables
- Methodology described in Foote, Machanavajjala and McKinney (JPC 2019)

United States® Census Bureau

Simulated Data



Adding noise




1
0.75
0.5
0.25
0
Cumulative Density
10
20
30
40
50
60
70
80
90
100
150
Earnings
True CDF


1
0.75
0.5
0.25
0
Cumulative Density
10
20
30
40
50
60
70
80
90
100
150
Earnings
True CDF
Protected CDF

Advantages of this DP implementation:

- We can aggregate to higher levels and calculate percentiles at those levels (for instance, median earnings at an institution)
- Easily implemented in other situations – Veterans Employment Outcomes used the same method, but with different histogram cutoffs
- Protection is intuitive for general public
- Code available for broader use (public on GitHub)

United States® Census Bureau

Table 1: Errors in Median Earnings, by Cell Size

Cell Size	Mean Absolute L1 Error	Mean Relative Error	L1 Error	
			10th Percentile	90th Percentile
30-50	2596	0.04714	-3584	4054
50-80	1526	0.0273	-2164	2344
80-100	1096	0.02108	-1642	1683
100-200	783	0.01461	-1193	1191
200-300	568	0.00925	-844	852
300+	344	0.00524	-523	506

Notes: Authors calculations.

Source: Foote, "Measuring Protection-Induced Errors in Earnings Outcomes from PSEO" (2021)

United States® Census Bureau

Questions?

andrew.foote@census.gov

Code Here

United States® Census Bureau

Extra Slides

PSEO Output Tables

- **Graduate Earnings Table**
  - 25^{th}, 50^{th}, and 75^{th} percentiles of annual earnings for college and university graduates
  - By degree level, degree major, and post-secondary institution
  - One year, five years, and 10 years after graduation
- **Employment Flows Table**
  - Employment by industry sector and Census Division of the country
  - By degree level, degree major, and post-secondary institution
  - One year, five years, and 10 years after graduation

United States® Census Bureau

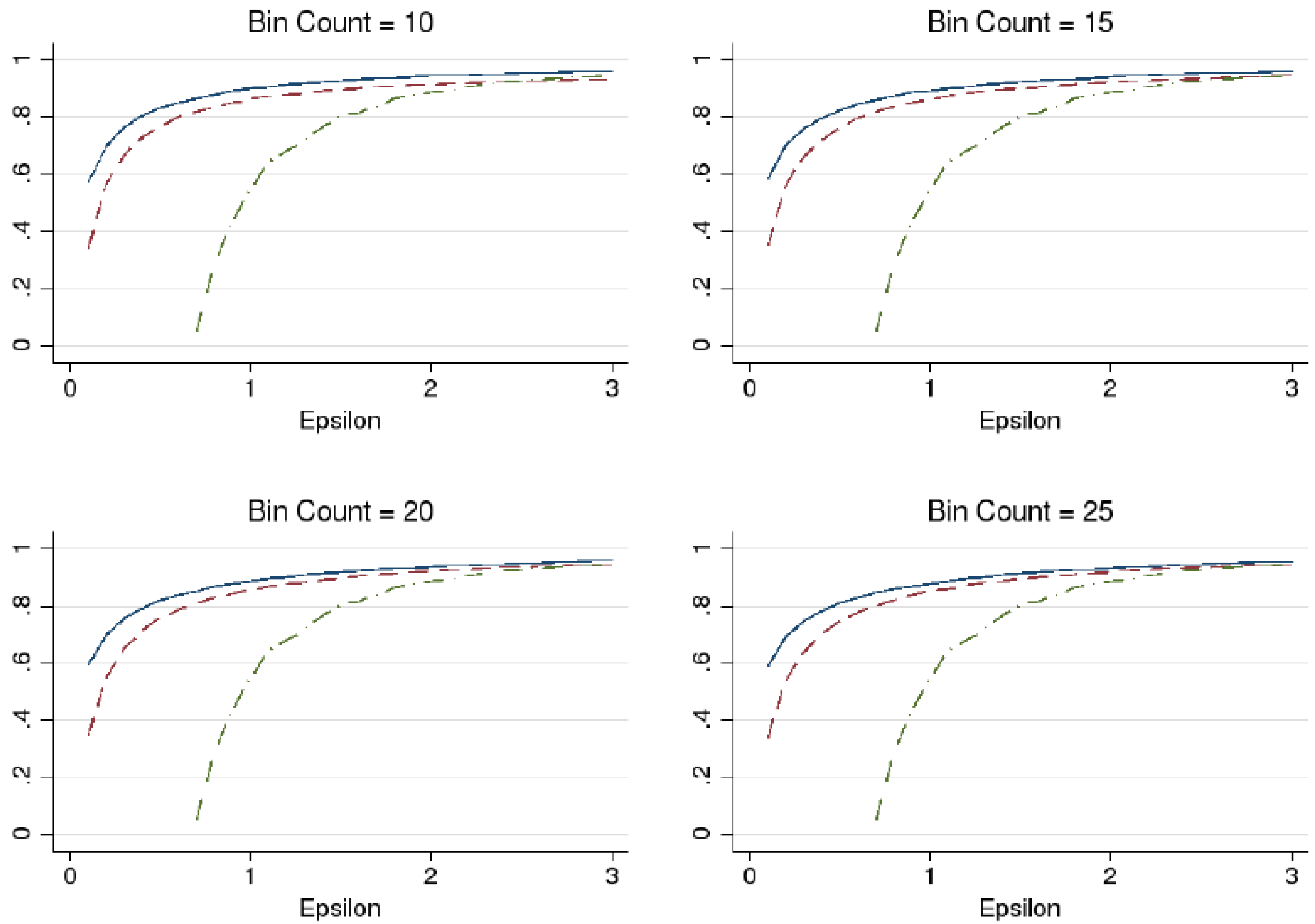
How well does this protection system perform?

- We compare this protection system to two other candidates:
  - Evenly spaced bins
  - Smooth sensitivity (Nissim et al 2007)
- We measure relative accuracy:

$$RelAccuracy = 1 - \frac{|Protected - True|}{True}$$

- Choices for implementer: privacy loss; bin count

Figure 1. Relative Accuracy by Histogram Method, 50th Percentile



Source: Foote, Machanavjjhala and McKinney (JPC 2019).
Blue is log-normal histogram; red-dash is even histogram; green dash-dot is smooth sensitivity